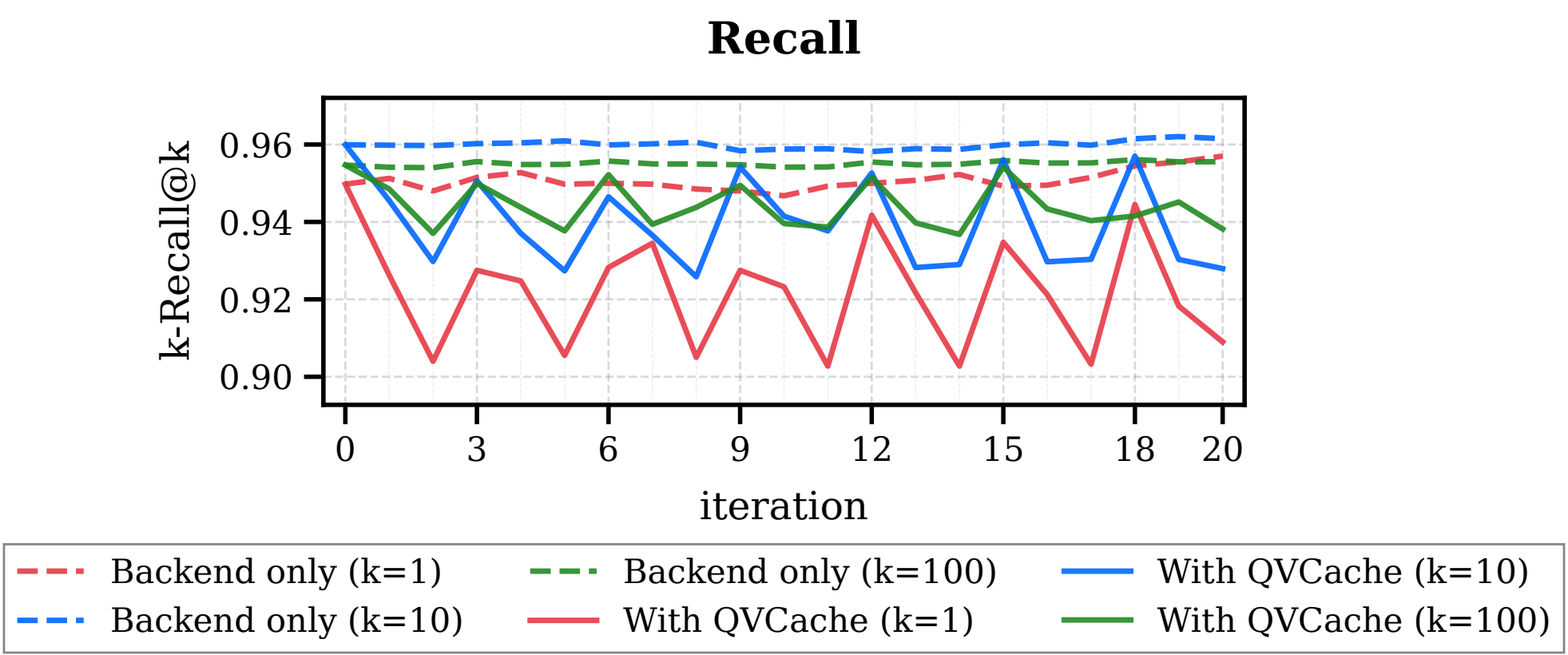
Recall
k-Recall@k
0.96
0.94
0.92
0.90
0
3
6
9
12
15
18
20
iteration
Backend only (k=1)
Backend only (k=10)
Backend only (k=100)
With QVCache (k=1)
With QVCache (k=10)
With QVCache (k=100)